GNCT Research Proposal

Geraldi Dzakwan – 13514065

Bandung Institute of Technology

April 21, 2017

1. Research Title

Voice Anonymization in Mobile Platform using Sensitive Named Entity Recognition, Vocoder, and Pitch Modifier

2. Research Background

There are so many conversations/interviews which shouldn't be disclosed to third parties as the dialog contains personal information about the speaker and/or characterizes the speaker. An example of this would be when a patient is telling about their medical condition to a doctor. This data (the speech) could be used for a further purpose. For example, it can be used for research/analysis purposes by a medical institution. Because privacy of personal information is very important, the speech should somehow be anonymized. In brief, voice/speech anonymization is an activity to mask all the sensitive data and to reduce the voice characteristics as much as possible (make it unrecognizeable) from a user's speech but in a way that do not remove the required data for research (e.g. symptoms explanation). Hence, it will minimize the risk of speaker's identity being discovered so the personal data won't be misused. Hopefully, the product outcome of this research can solve the initial problem and can be useful for extensive range of user segments (e.g. patient, customer, and witness).

3. Research Purpose

The purpose of this research is to create a mobile platform that could convert a speech to it's anonymous version so that the speech could be published for the sake of research, analysis, etc. The outcome of this research will be deployed in mobile platform. This mobile platform aims to record audio data and then deliver it without disclosing any private information and without exposing the voice characteristics.

This system could be used for several purposes, such as :

1. Analyzing anonymous medical interviews (by not revealing patient's identity) for the sake of research in local medical institution

2. Analyzing complaints for sentiment analysis without disclosing customer's identity
3. Anonymize the speech of witness in court so the witness can talk freely without any pressure on him/her

4. Research Scope

In this research, the speech data will only be in English. The speech data should have clear pronunciation (the speech will be recorded in a very silent place). The system can provide the output as either voice or text but it can't accept text input.

5. State of the Art (Literature Study)

There are several well-known projects and publications related to anonymization that I used as references which are :

1. MITRE, Wellner *et al*., 2006
2. Szarvas *et al*. System, 2006
3. Arakami *et al*. System, 2006
4. HIDE, Gardner *et al*., 2008
5. Jianfeng Chen. Dat Tran Huy, et al., " Using Keyword Spotting and Replacement for Speech Anonymization", 2007
6. Henning Pätzold, "Secondary Analysis of Audio Data. Technical Procedures for Virtual Anonymisation and Modification", 2005

Paper number 1-4 is about how to determine the sensitive named entity from given sequence of words. In this proposal, the proposed identifier method comes from combining the method in paper 2 and 3 and add my own idea/method to it. Meanwhile, paper number 5 and 6 is about the anonymization techniques commonly used for speech data. Paper 5 contains quite a simple way to identify the sensitive named entity using only one model. So, I try to improve the process by implementing the proposed identifier method (with paper 1-4 as references). Last, I try to implement vocoder and pitch changer/modifier explained in paper 6.

6. Proposed Method

This research will try a modified method on how to do voice anonymization for English voice. In this research, I will use common speech which contains unstructured data represented in natural language in English. Because the it is a free-form speech, then we need to identify each word that represents unique identifier. This is later be called as “entitiy” in this proposal. Most of anonymization systems are built by implementing these four steps :

1. Normalize word and extract feature
2. Build Name Entity classifier (if more than one can be paralleled)
3. Decide the class of the entity (most probable one)
4. Applying anonymization method for every entity occurrence (e.g. replacing them with suppression)

6.1. NLP Tools Used

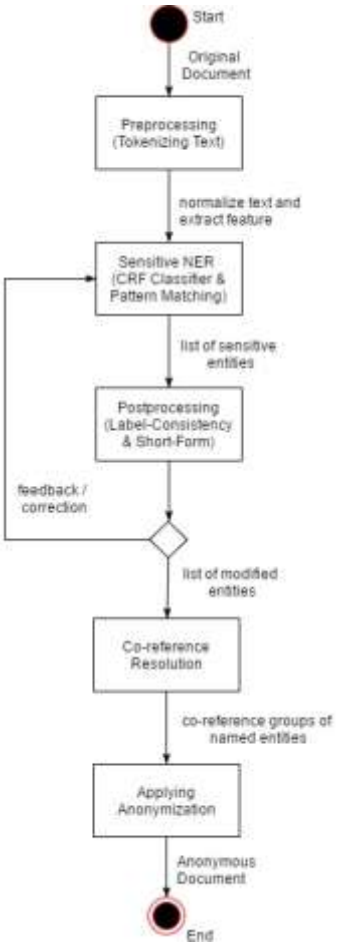
There are several tools that can be used for the natural language processing, such as Stanford’s CoreNLP and Natural Language Toolkit. They both have their own strengths and weaknesses. But, I am going to pick Natural Language Toolkit (NLTK) because I am going to use Python for this project (as it is my first choice of programming language).

6.2. Speech to Text Translation

I am going to use tools like CMU Sphinx or other tools that use Hidden Markov Model (like HLT) to accept speech input and translate it to text data which can be processed furthermore to determine the sensitive named entity.

6.3. Identifying Sensitive Named Entity and Doing the Anonymization

I proposed five steps to do the anonymization, which are below :



6.3.1 Preprocessing

This operation will be executed using the NLP tools mentioned before. The text will be normalized, brokedown into sentences, and it will end up as tokens. So, basically the key things here is to tokenize the text and give a clue for each token. The clue here can be a post tag or other morphosyntactic things (I am still working on this to determine what really should be done in this step). The clue later can be used for further process such as NER.

6.3.2 Sensitive Named Entity Recognition

This process will also be performed by the NLP tools because as far as I know they support this functionality. The reason are because NER tools provide a ready to use class entity recognition and the accuration is good based on previous works. This process will receive the tokenized text and return a list of entities that are found in the document. An entity will be an element of the list. It contains the information of it's position (for example it's the 200^{th} word in the document) and it's class (for example it's a name of a person). The classifier which will be used is CRF classifier along with

the pattern matching detection. For this research, I would likely to define six classes which are below :

1. Person
2. Location
3. Organization
4. Date
5. Time
6. Miscellaneous

But, I am a little bit not sure about the 4th and 5th. Because, usually only the first 3 classes supported by the tools. Another alternative is to classify NE to 4 classes only (Person, Location, Organization, Misc). To decide the most probable class, I will use the voting method.

6.3.3 Postprocessing

I haven't defined detail process for this step, but in general this step's purpose is to make sure the sensitive named entity recognition process does only few mistakes. In other words, to improve the performance. Two methods which will be used are short-forms and label-consistency method. This process will receive the list from Sensitive NER process and return the same list which has been modified first according to the correction applied.

6.3.4 Co-reference Resolution

This process will group all the named entities with the same object references. This is important since we want to preserve the context of original speech. If we don't perform this process, entities which refer to the same object would be replace by different expressions. Hence, the context of the speech will change and we don't want that. So, basically the purpose of this process is to replace all the entities referring to the same object by the same expression. I would likely to use the rule-based co-reference resolution. This process will receive the input from postprocessing step in the form of list and return the same list which has been modified. An entity will have one more attribute, which is the reference object.

6.3.5 Applying Anonymization

This process will replace all the entities based on the list it receives with expressions. It will add another attribute to the list which is a replacement expression. The replacement expression should be the same for each entity within the same group. After

the replacement expression has been added to the list, the words in the speech than be changed accordingly. There are several methods of replacement that I will use in this research which are :

1. Simply silence the voice. The system will take start time and end time of the word in speech and replace it with silence expression.

2. *Numbered tagging*
   This is a simple approach in which the replacement expression would be the class of the entity concatenated with an identifying number. For example, there are two names that will be replaced, say James and John in which James comes first in the text. Then, James would be replaced by "Person1" and "John" would be replaced by "Person2".

3. *Generalization*
   This approach is more complicated than the first approach. This approach will replace an entity by a more general entity but still in the same class. Let's take some examples. Say you have a word "Borromeus Hospital". Instead of throwing "Organization" as replacement expression, it would throw "Hospital". Another example is the word "Bandung". Instead of throwing "Location" as replacement expression, it would throw "City". Only the class "Person" that would not be generalized. This approach can be implemented by using WikiData. WikiData itself is a knowledge base in which you can determine the superclass of an entity. WikiData will find an entry whose name, pseudonym, or acronym matches the input entity. It then find the superclass of the entry until it matches the class of the entity (until it found "Location", "Organization", or "Miscellaneous"). For this research, I would likely to choose a class which is a level or two level below the superclass to maintain consistency.

6.3.6 Applying Vocoder

Using this technique, the result is an audio data in which the voice modulation of the speaker is transferred to a white noise, so that the voice is no longer recognizable (the sex of the speaker is also unidentifiable). The further switches of the program usually work well with the defaults. But, there are some problems if this technique is used. The most problematic aspect of this procedure is the loss of audio quality in the output speech. As long as the source was recorded carefully it is still reasonable, though understanding the speakers becomes a little more difficult, and it still keeps open all possibilities for further processing. Hesitations and breaks remain unaltered, stresses and nonverbal sounds are also still recognizable, though of less intensity. If, however, the source file is of poor quality, the resulting file may not be wholly comprehensible and analyzable.

Furthermore, the noise can replace the speaker's voice. Therefore, a mapping of statements to speakers by voice is no longer possible. In the case of interviews with one person, it is helpful to have a stereo recording which allows one to separate the tracks and alter only that of the interviewee (or both with different parameters).

6.3.7 Applying Pitch Modifier

Pitch changing moves the whole spectrum of frequencies up or down (voices in the modified audio file are therefore higher or lower than in the original). The difference in pitch between the different voices and other characteristics are preserved with greater authenticity than with the vocoder. Convincing modification which preserves good comprehensibility is achieved by changing the pitch to a middle range which means raising low voices and lowering high ones. Good results are often achieved with a variation of about 3-5 half tone steps. Of course, it is always necessary to compromise if the voices differ greatly in pitch because the anonymization of the interviewee is the most important goal.

Hence, there are problems if pitch modifier is used. Anonymization using pitch modifier is weaker because more characteristics of the voice are kept; at the same time it makes further analysis easier. An audio file of bad quality, which would be totally unintelligible using the vocoder, might still be usable after pitch shifting. The deciding factor is the amount of anonymization required. Voices which differ markedly in pitch can cause problems for pitch changing. In such cases, much like the situation described above, a stereo recording is helpful as it allows one to modify different tracks in opposite directions. Furthermore, in principle this allows only the voice of the interviewee to be changed, while leaving the interviewer's unaltered.

Thus, because vocoder and pitch modifier techniques have their respective weakness, I try to use a deciding factor regarding which technique should be used based on the speech input. Example parameter to be considered is the voice quality.